

# Developing embedding models for Scottish Gaelic



# Developing Word Embedding Models for Scottish Gaelic

William Lamb[1] Mark Sinclair[2]
(1) Celtic and Scottish Studies, University of Edinburgh, School of Literatures, Languages and Cultures, 50 George Square, Edinburgh EH8 9LH, United Kingdom
(2) The Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom
w.lamb@ed.ac.uk, mark.sinclair@ed.ac.uk

## RÉSUMÉ

**Développement de modèles vectoriels continus de mots pour le gaélique écossais**

Nous présentons ici un projet préliminaire pour la construction et l'évaluation de représentations vectorielles continues des mots appliquées au Gaélique écossais. Les méthodes de représentation vectorielles continues des mots ont déjà été appliquées avec succès à de nombreuses tâches en traitement automatique de la langue (TAL) et ont pour avantage de pouvoir être construites à partir de texte brut et non structuré. Ces méthodes sont ainsi particulièrement adaptées aux langues faiblement dotées en ressources linguistiques telles que le Gaélique. Nous avons construit trois différents modèles vectoriels continus des mots à partir de deux versions d'un corpus de 5.8 millions d'occurrences de mots (tokens). La première version consiste contient la simple segmentation en occurrences alors que la deuxième version comprend les occurrences et les formes lemmatisées. La représentation syntaxique des modèles est évaluée à partir d'un étiqueteur syntaxique en Part-of-Speech (POS). Par ailleurs, diverses requêtes sémantiques effectuées sur les modèles permettent de mesurer et caractériser leur richesse sémantique. Les modèles construits à partir du corpus d'occurrences seules s'avèrent peu robustes aux requêtes sémantiques en raison de la parcimonie des données. En revanche la lemmatisation améliore la robustesse des modèles pour les requêtes sémantiques mais au prix d'une sensibilité flexionnelle accrue. Nous illustrons les différences entre les modèles ainsi que l'apparent compromis entre leurs capacités sémantiques et syntaxiques. Finalement, nous soulignons le potentiel des représentations vectorielles continues des mots pour toute une série d'applications futures.

## ABSTRACT

**Developing Word Embedding Models for Scottish Gaelic**

We detail a preliminary project on encoding and evaluating word embeddings for Scottish Gaelic. Word embedding methodologies show promise for diverse natural language processing (NLP) tasks and can be built from raw, unstructured text. Accordingly, they are attractive for under-resourced languages like Gaelic. We instantiated three embedding models on two versions of a 5.8 million token corpus : 1) tokenised and 2) tokenised / lemmatised. Using a simple POS tagger, we quantitatively measured the syntactic similarity between nearest neighbours for each model's vector-space representations of words. We also queried the models to assess their semantic specificity and breadth. Models built from the tokenised corpus exhibited the effects of data sparsity for semantically constrained queries. The lemmatised versions had more semantic robustness, but at the expense of inflectional sensitivity. We note divergences between the models and an apparent inverse relationship between their semantic and syntactic capacities. Finally, we highlight the promise of word embeddings for a range of future work and downstream applications.

MOTS-CLÉS : gaélique écossais, modèles vectoriels continus de mots.

KEYWORDS: Scottish Gaelic, word embeddings, neural networks, natural language processing, word2vec, part-of-speech tagging.

# 1 Introduction

When reflecting on the position of our language technologies, it is common for those working on minority languages to express some degree of English envy. State-of-the-art natural language processing (NLP) technologies typically require large quantities of labelled training data. These are readily available for English and other majority languages, but not normally for under-resourced languages. Yet, as in other data-driven fields, NLP has recently been dominated by approaches leveraging artificial neural networks. While these approaches do not necessarily mitigate requirements for labelled data directly, they are attractive for their language-independence and the fact that they can be generated unsupervised from relatively raw data (Lin *et al.*, 2015; Chen *et al.*, 2013). Large annotated corpora are unlikely to exist for under-resourced languages, but copious amounts of on-line text are often available. After light processing, this text can be made suitable for approaches based on neural networks. As we demonstrate in this paper, useful models can result from modestly-sized datasets.

A key difference between a neural network and conventional NLP approach is that the former typically requires words to be represented as numerical vectors. The process of mapping atomic word units (tokens, lemmas, etc.) to vectors is known as 'word embedding'. Neural network word embeddings, or vector space models (VSMs), use high-dimensional geometry to map associations between words. Embedding algorithms exploit the iconic relationship between semantics and linguistic context, typically mapping similar words to nearby vector points. The underlying principal recalls Firth's observation that '[y]ou shall know a word by the company it keeps' (1957 : 11). Although vectors are difficult to interpret — each dimension represents multitudinous concepts and concepts are spread multi-dimensionally (Al-Rfou *et al.*, 2013) — word embedding models have proven effective as input to a variety of standard NLP tasks, such as part-of-speech (POS) tagging (Fonseca *et al.*, 2015).

Given the above characteristics and possibilities, word embedding models could be useful for work involving Scottish Gaelic (Al-Rfou *et al.*, 2013). Although improvements have been made to Gaelic language technology in recent years (see Batchelor, this volume), it still lags behind that of most larger languages and even some minority languages (e.g. Irish Gaelic). We present this paper as proof of concept in the interest of using word embeddings methodologies to expedite the development of Scottish Gaelic NLP resources. In the sections below, we overview our methodology, provide an initial assessment of the models' strengths and weaknesses and comment on potential downstream applications and future possibilities.

# 2 Background and Methodology

Scottish Gaelic is a Celtic language that is closely related to Irish and Manx Gaelic, and more distantly related to Cornish, Welsh and Breton. Working with Gaelic in an NLP context presents several challenges. As aforementioned, one is the low availability of high quality data, such as tagged

corpora. [1] Additionally, word forms in Gaelic are remarkably protean due to its complex morphology (see Lamb, 2008). For example, its nominal system features initial and terminal mutations that are sensitive to grammatical categories such as case, number and definiteness. A word like *cailleach* 'old woman' may appear as *chailleach, caillich, cailliche, chailliche, cailleachan, chailleachan, cailleachaibh* or *chailleachaibh*, depending on grammatical context. (From the lexicon described below, we calculate an average surface-form to lemma ratio of 6.84 to 1.) In addition, although orthographic standards exist (SQA, 2009), few writers adhere to them exclusively ; spelling can be idiosyncratic. Given this variability, data sparsity is a common problem as we discuss further below.

Our data came from a 21 million word web-crawl of Gaelic text, available as part of the Crúbadán [2] project (Scannell, 2007). The source texts are diverse in register and quality, ranging from biblical prose to chat room dialogue. Much of the text stems from the Gaelic version of Wikipedia (gd.wikipedia.org). Scannell took a random sample of sentences from larger sources to lessen any bias towards them, and extirpated the data of much ambient English. He provided us with a file of 5.8 million tokens (263,858 lines ; 133,287 unique tokens) and, from this, we generated two training files : 1) tokenised and 2) tokenised and lemmatised. We built the lemmatiser [3] using a large, manually constructed lexicon of Scottish Gaelic (Am Faclair Beag : [4] see Patton, 2016). It is capable of handling common orthographic variations, such as 'Uidhist' for 'Uibhist' (Eng : 'Uist'), although not out-of-dictionary items ; for this study, we replaced the latter with "#IGNORE". Light standardisation was applied in the form of automatically de-capitalising wrongly capitalised tokens.

## 2.1 Word Embedding Method

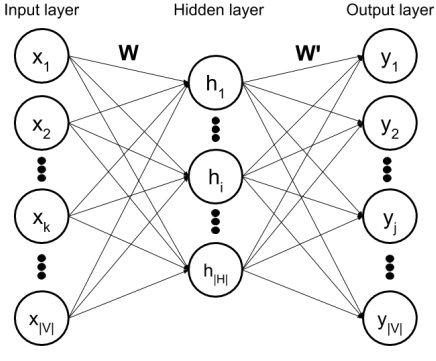


FIGURE 1 – A typical multi-layer perceptron used for learning vector-space word embeddings.

Figure 1 shows a typical multi-layer perceptron (MLP) with input, hidden and output layers. The input and output layers have a node for every word in a given vocabulary $V$, of size $|V|$. Each of these layers is fully connected to a single hidden layer of size $|H|$. The connections are represented by input and output weight matrices, $W$ and $W'$, respectively.

1. But see (Maolalaigh, 2013; Lamb *et al.*, 2016)
2. *Crúbadán* is Irish for 'crawler' : see http ://crubadan.org
3. We hope to develop the lemmatiser further and make it available in the future
4. www.faclair.com

The nodes in the hidden and output layers perform simple functions that aggregate their inputs and produce a single output. These functions are generally fixed to be common across all nodes in a given layer (specifically they are often softmax or sigmoid functions). It is the weight matrices that describe how to emphasize or de-emphasize a given input to a node. By learning the weights that maximise the likelihood of input/output example pairs, the network can learn inherent structures within a given dataset. Once the learned weights are fixed, we can present an arbitrary input vector to the network and compute a corresponding output vector.

As an example, consider the following dataset of word pairs (bi-grams) that describe Scottish geographical features — lochs (lakes), rivers and bens (mountains) — along with specific names.

$$D = \{loch|lomond, river|ness, ben|lomond, loch|ness, river|tay, ben|vorlich, loch|tay,$$
$$river|clyde, ben|more, loch|more, river|forth, ben|nevis\}$$

We may wish to have the network learn to associate a feature with a name or vice-versa. The vocabulary required to describe this complete set would be :

$$V = \{ben, clyde, forth, loch, lomond, more, ness, nevis, river, tay, vorlich\}$$

Each pair of words can then be described as an input and output vector that is 1 at the position of the word and 0 otherwise , e.g. for the pair $loch|lomond$ :

$$loch = \{0,0,0,1,0,0,0,0,0,0\}, lomond = \{0,0,0,0,1,0,0,0,0,0\}$$

By providing the network with these "one-hot" vectors during training, it can learn the weight matrices (typically by means of the back-propagation algorithm) that are best able to make a correct mapping from input to output. As the number of hidden units $|H|$ is typically much smaller than $|V|$, it compresses the input through the hidden layer and decompresses it at the output. It is these compressed vectors that allow us to encode words into a more condensed vector space than that of the input or output layers. We can then measure the distance between vectors in order to find out how 'close' words are in a given model. In this example, we may expect to find $clyde$ close to $river$ and far from $ben$ and $loch$ because there are no such places. However, $lomond$ may be close to both $ben$ and $loch$ as both places exist, so it can co-occur with either of these words. This is a very simple example but serves to show how such neural network architectures can be used to perform NLP tasks such as answering the question "which word comes next ?". Derivatives of similar architectures can be used to model more complex relationships betweens words.

## 2.2 Tools and Model Types

In order to train the word embedding models, we used the tool [5] developed by (Ling *et al.*, 2015), which itself is a modified version of the popular word2vec [6] algorithm (Mikolov *et al.*, 2013a,b,c). This tool allows word vector representations to be learned from any raw text corpora. Several different learning schemata have been made available so we chose three of the most popular, as described briefly below.

**Constrained Bag-of-Words (CBOW)**

This training schema works by learning to predict a word given a context. For example, if we consider that we have 5-gram training examples such as “quick brown fox jumps over”, we can set our input vector to a one-hot representation of “quick brown jumps over” and the output vector to “fox”. The model will then learn weight matrices that can predict this

5. https ://github.com/wlin12/wang2vec
6. https ://code.google.com/archive/p/word2vec

relationship.

**Skipngram**
Skip-gram works as a kind of 'inverse' of *CBOW* – given an input word, we want to predict the context. The step size for the N-grams used for context can also be altered to provide a wider context.

**Structured skip-gram**
Structured skip-gram works in a similar way to the standard skip-gram model, except that we also provide the relative positions of each context word with respect to the target word. This allows the model to learn more about the inherent linguistic structure such as the proximity of adverbs to verbs or adjectives to nouns. As a consequence, the model can potentially better learn syntactic relationships between words.

The Polyglot project [7] (Al-Rfou *et al.*, 2013) offers a selection of word embedding models that have been automatically generated for a number of languages, including Scottish Gaelic. While this model was trained on different data, we were still able to use it for comparisons.

Unless otherwise stated, all of our models were trained with a 5-gram window over 3 iterations of training with a hidden layer size of 64 (to match Polyglot) or 200. Any words with a frequency of less than 5 were ignored.

# 3 Evaluation

## 3.1 Overview

The evaluation of word embedding models is typically performed in one of the following three ways :

**The use of a lexical relationship database**
For some well resourced languages such as English, the availability of a well-curated lexical database of word relationships can be exploited to analyse word embedding models. WordNet (Miller, 1995) is an example of such a database. In WordNet, words are divided into high level syntactic classes (noun, verb, adjective, adverb, etc.) which are then grouped into sets of cognitive synonyms (synsets). Within each synset, words are interlinked according to conceptual-semantic and lexical relations. Given such a resource, it is possible to evaluate if a word embedding model is in agreement with the database across many different criteria. We can, for example, query the database for relations of a given word and then check the cosine distance between the word and each relation in the embedding model — whereby lower distance would indicate better agreement with the database (Handler, 2014). Lexical databases such as WordNet are the product of a substantial cumulative effort involving thousands of work-hours from expert annotators and, unfortunately, a similar resource does not yet exist for Scottish Gaelic.

**Subjective experiments**
The concept of word similarity inherently contains a subjective component, particularly concerning semantic relationships. Therefore, it is pertinent to consider designing subjective experiments to evaluate word embedding models. Typical forms of such experiments may include asking subjects to : rate or rank word groups in order of similarity ; suggest a similar

7. https ://sites.google.com/site/rmyeid/projects/polyglot

word y given word x ; choose a best match or selection preference, e.g. which noun typically goes with this verb ? Such queries can also be posed to the word embedding models so that answers can be compared with human subjects.

There are well-documented issues associated with subjective evaluations such as the above, one being inadequate sample size. The recent trend of using crowdsourcing technologies such as Amazon Mechanical Turk (AMT) provides the opportunity to alleviate some of these issues (Schnabel *et al.*, 2015). AMT allows users to set up web-based experiments whereby volunteers can participate for a small financial reward. However, this requires a large pool of potential subjects in order to yield enough respondents. We feel it would be unlikely to garner a large enough response from Scottish Gaelic speakers by means of AMT due to its relatively low number of native speakers. However, in the future we may consider smaller scale subjective experiments that do not utilise crowdsourcing.

**Downstream task performance**

Instead of evaluating the performance of word embedding models directly, we can also evaluate how they affect other downstream NLP tasks. For example, word embeddings from a given model could be used to train a language model or a document classifier. We can then use evaluation metrics and/or datasets for those tasks. This can often be a more informative evaluation if the word embedding model is designed for a specific task.

## 3.2 Syntactic (quantitative)

We were able to derive a quantitative measure of syntactic representation for each model by exploiting an existing, manually composed lexicon (see Patton, 2016). This lexicon provides highly detailed information with respect to potential POS tags for a given token, including all valid combinations of case, gender, number, etc. In total, the lexicon offers 198 possible POS tags, reflecting the relatively rich morphology of Scottish Gaelic as compared with languages such as English (the Penn Treebank Project, for example, considers only 36 POS tags).

We do not expect our models to able to distinguish accurately between POS tags at a high level of granularity, so we considered only the first-order tags of the lexicon : verb, noun and adjective. We then removed all tokens from each model that did not have a corresponding entry in the lexicon. This meant that we were able to look at the $n$-nearest neighbours of a given token within the vector space of each model and observe any homogeneity among the associated POS tags. By counting how often the POS tags of a token and its neighbours are in agreement, we are able to quantify the tendency of each model towards a more syntactically informed clustering.

The syntactic score, $S$, is formulated as follows :-

$$S = \frac{\sum_{i=1}^{|V|} \sum_{j=1}^{n} f(POS_i, POS_{i,j})}{|V|n}, f(A,B) = \begin{cases} 1 & \text{if } a = b \text{ for any } a \in A, b \in B \\ 0 & \text{otherwise} \end{cases}$$

Here, $POS_i$ represents the set of part-of-speech tags for word $i$, and $POS_{i,j}$ represents the set of tags for the $j$-th nearest neighbour to word $i$ according to cosine distance.

Table 1 shows the results for several model types. In order to compare with the Polyglot model, we used the 2000 most frequent tokens that were in common across all models to calculate the score. We present results with a hidden-layer size of $|H| = 64$, which matches the Polyglot model, and also with a larger value of $|H| = 200$. In all cases the smaller hidden-layer performs better. This may be an effect of inherent regularization offered by having fewer parameters to model our dataset which is

TABLE 1 – Agreement of tokens with N-nearest vector neighbours in POS category (Syntactic Score)

| | **Syntactic Score,** $S$ | |
|---|---|---|
| **Model** | $\|H\| = 64$ | $\|H\| = 200$ |
| Polyglot | 64.87% | - |
| CBOW | 82.11% | 82.08% |
| skipgram | 75.06% | 73.59% |
| structskipgram | 85.06% | 84.32% |

still relatively small for this task. If we had more data, then a larger hidden-layer may perform better.

We find that *CBOW* performs better than *skipgram*, which is consistent with other findings in the literature (Mikolov *et al.*, 2013a; Qiu *et al.*, 2014). As expected, adding structural information to the *structskipgram* model significantly increases the syntactic score. This is likely due to the extra information helping the model to learn local grammatical structures that can push words into certain clusters based on their relative positions. The Polyglot model has the lowest performance. This may be due to the differences in training data or it could be simply that the model was designed to capture another type of linguistic relationship. Due to the near-fully automatic method used for training the minority Polyglot languages the training data may only have gone through generic — rather than language specific — tokenisation.

It is worth noting, however, that improvement in syntactic modelling may be at the expense of semantic modelling and a suitable choice of model may depend on the target application. A strong syntactic model may, in future work, be able to provide supplementary information to tasks such as part-of-speech tagging for Scottish Gaelic.

## 3.3 Semantic (qualitative)

As expected from a base of 5.8 million words, the models capture robust semantic and syntactic relationships between common words. However, they lack the sophisticated nuances reported for languages with more available text. We begin by discussing the differences between the models, followed by how well the models encode semantic information. It is worth noting that the empirical evaluation of word embeddings semantics is at an early stage. Although paradigms exist, as detailed above (Schnabel *et al.*, 2015), they require significant human resources. Our comments, perforce, are impressionistic at this point.

We queried the models with terms selected to test their semantic granularity and breadth. Although the models provide similar returns for very common words, they diverge notably with more semantically constrained ones. For example, in Table 2, we report the top five returns ranked by cosine similarity for *Uibhist* 'Uist', a well-known Hebridean island. (NB : Unless noted with '*', the models were trained on the lemmatised data.)

From this result and others, it would appear that the models are sensitive to different semantic domains. *CBOW* is effective at locating the general semantic category ; it groups 'Uist' with a variety of other place-names. *Skipgram* returns nearby island place-names only, indicating greater specificity. *Structskipgram* is similar to *skipgram*, but includes 'America'. When considering returns for *eaglais* 'church', we see similar tendencies : *CBOW* returns other buildings (e.g. hotel, palace, abbey) while *skipgram* returns other ecclesiastic nouns (parish, Catholic, abbey, graveyard). Again, *structskipgram*

TABLE 2 – Nearest neighbours per model for input query Uibhist 'Uist' (lemmatised). English translations are provided for convenience.

| **CBOW** | **skipgram** | **structskipgram** |
|---|---|---|
| *Aimeireaga* 'America' | *Èirisgeigh* 'Eriskay' | *Tiriodh*'Tiree' |
| *Èirisgeigh* 'Eriskay' | *Barraigh* 'Barra' | *Ìle* 'Islay' |
| *Afraga* 'Africa' | *Tiriodh* 'Tiree' | *Slèite* 'Sleat' |
| *Barraigh* 'Barra' | *Slèite* 'Sleat' | *Leòdhas* 'Lewis' |
| *Àisia* 'Asia' | *Leòdhas* 'Lewis' | *Aimeireaga* 'America' |

TABLE 3 – Nearest five neighbours for common semantic domains. English translations are provided for convenience.

| **TOKEN** | **TRANS.** | **TOKEN** | **TRANS.** | **TOKEN** | **TRANS.** |
|---|---|---|---|---|---|
| ***mara**** | **sea (gen)*** | ***obh*** | **oh (dear)** | **faicinn** | ***seeing*** |
| *beinne* | hill (gen) | *Obh* | Oh (dear) | *cluinntinn* | hearing |
| *coille* | forrest (gen) | *siuthad* | go on | *tuigsinn* | understanding |
| *gaoithe* | wind (gen) | *och* | oh | *faireachdain* | feeling |
| *mòintich* | moor (gen) | *ist* | listen | *saoilsinn* | thinking |
| *creige* | rock (gen) | *siuthadaibh* | go on (pl) | *smuaintinn* | thinking |
| ***dearg*** | ***red*** | ***beagan*** | ***a bit*** | ***bus*** | ***bus*** |
| *geal* | white | *cus* | too many | *bàta* | boat |
| *gorm* | blue | *tòrr* | many | *trama* | tram |
| *glas* | gray | *mòran* | many | *trèana* | train |
| *uaine* | green | *barrachd* | more | *trèan* | train |
| *donn* | brown | *moran* (sic) | many | *plèan* | plane |
| ***(An) Eadailt*** | **Italy** | ***dithis*** | **two people** | ***craobhan**** | **trees*** |
| *(An) Ruis* | Russia | *triùir* | three people | *creagan* | rocks |
| *(An) Ostair* | Austria | *ceathrar* | four people | *glinn* | glens |
| *(An) Fhraing* | France | *dithist* | two people | *lusan* | plants |
| *(An) Òlaind* | Holland | *còignear* | five people | *cnuic* | hills |
| *(An) Spàinn* | Spain | *sianar* | six people | *rathaidean* | roads |

is intermediate in focus. Unless otherwise stated below, we report results from *CBOW*, which seemed to be the most semantically coherent model overall.

As reported in Table 3, the model discriminates common semantic domains such as colours, countries and modes of transport. [8] Interestingly, it also groups returns based upon the case, number and grammatical category of the query word. For example, in the case of *obh* 'oh (dear)', the model returns other interjections. The genitive of the feminine noun *muir* 'sea' prompts other feminine, genitive nouns associated with physical geography. Quantifiers and quantitative pronouns are also grouped together, as are psychological verbal-nouns (e.g. *creidsinn* 'believing'). When queried, *craobhan* 'trees' produces other landscape-oriented plural nouns.

These results are promising given the relative paucity of data and the sparsity associated with Gaelic morphology (see Danso & Lamb, 2014). However, these issues come into sharp relief with other

8. Note : *trèana* is a variant of *trèan* as *dithist* is a variant of *dithis*

TABLE 4 – Effects of data sparsity : tokenised vs lemmatised models for clàrsach 'harp' (structskipn-gram reported)

| **TOKEN** | **TRANS.** | **LEMMA** | **TRANS.** |
|---|---|---|---|
| ***clàrsach*** | **harp** | ***clàrsach*** | **harp** |
| *teip* | tape | *pìob* | bagpipe |
| *coimpiutairean* | computers | *druma* | drum |
| *dubhan* | hook | *fonn* | tune |
| *fònaichean-làimhe* | mobile phone | *giotàr* | guitar |
| *tulach* | hill | *seinn* | singing |

queries. For example, the noun *clàrsach* 'harp' – when run through the models trained on the tokenised corpus – produces seemingly unrelated nouns (see Table 4). However, when queried on the models instantiated from the lemmatised corpus, other musical instruments and terms are returned as expected. As is well known, lemmatisation is an effective way to handle data sparsity. On the other hand, it restricts potential searches to root forms and this can be disadvantageous when grammatically sensitive models are required. Additionally, lemmatisers can introduce their own errors, and exclude a significant proportion of tokens as 'out of vocabulary' if the data is orthographically inconsistent. Therefore, one must balance potential gains and losses when considering whether to use them.

# 4 Conclusions

Although the word embedding techniques employed here are already well-represented in NLP literature, this is the first example of their application and evaluation for Scottish Gaelic. We adapted existing resources for our task, and trained and evaluated a variety of models on two versions of the textual data : 1) tokenised and 2) tokenised and lemmatised. Although the resources required for conventional evaluation approaches were not available, we were able to derive an objective measure of syntactic modeling capacity and an initial, qualitative assessment of semantic modeling capacity. We find the performance in each category to be dependent upon the choice of the model, potentially with an inverse relationship obtaining between the two. In other words, improving syntactic performance may be at the expense of semantic performance, although additional work is required.

The relative lack of resources for Scottish Gaelic compounded with the language's morphological complexity presents significant data sparsity issues. We have shown how these issues can be partially mitigated by lemmatising the training data *a priori*. However, the lemmatisation process introduces its own errors into the overall end-to-end system and may not be suitable for applications requiring sensitivity to grammatical inflection.

A motivating factor for this study, as aforementioned, is that approaches based upon word embeddings facilitate the exploitation of raw textual data without the need for manual annotation or intervention. Therefore, they provide a gateway for dealing with the resource constraints that commonly face minority languages.

We anticipate applying what we have learned from this study in two ways : 1) improving the model training by bootstrapping from better-resourced, related languages such as Irish and Welsh (e.g. by pre-training the model on those languages before fine-training on Scottish Gaelic) and 2) substituting conventional atomic representations with vector-space representations for a variety of potential

downstream NLP tasks (e.g. POS tagging, language modelling and machine translation). Vector space representations are a prerequisite for accessing artificial neural network solutions, which are increasingly driving state-of-the-art language technology. Therefore, this initial work is promising for the future of Gaelic NLP.

## Acknowledgements

We wish to thank Prof Kevin Scannell (Saint Louis University) for providing the data that we used to train the models. We are also indebted to Michael Bauer and Will Robertson of 'Am Faclair Beag' [9] for allowing us to use their lexical database.

## Références

AL-RFOU R., PEROZZI B. & SKIENA S. (2013). Polyglot : Distributed word representations for multilingual NLP. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, p. 183–192.

ANDREAS J. & KLEIN D. (2014). How much do word embeddings encode about syntax ? *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, p. 1–9.

BATCHELOR C. (2016). Automatic derivation of categorial grammar from a part-of-speech-tagged corpus in Scottish Gaelic. In *Proceedings of the Celtic Technology Workshop (CLTW 2016)*, volume 2.

CHEN Y., PEROZZI B., AL-RFOU R. & SKIENA S. (2013). The expressive power of word embeddings. In *Proceedings of the 30th International Conference on Machine Learning*, p. 1–9.

DANSO S. & LAMB W. (2014). Developing an automatic part-of-speech tagger for Scottish Gaelic. In J. JOHN, L. THERESA, W. MONICA & B. Ó RAGHALLAIGH , Eds., *Proceedings of the Celtic Technology Workshop (CLTW 2014)*, volume 1, p. 1–5.

FIRTH J. R. (1957). *A Synopsis of Linguistic Theory, 1930-1955*. Blackwell.

FONSECA E. R., ROSA J. L. G. & ALUÍSIO S. M. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, **21**(1), 1–14.

HANDLER A. (2014). *An empirical study of semantic similarity in WordNet and Word2Vec*. PhD thesis, Columbia University.

LAMB W. (2008). *Scottish Gaelic Speech and Writing : Register Variation in an Endangered Language*, volume 16 of *Belfast Studies in Language, Culture and Politics*. Cló Ollscoil na Banríona.

LAMB W., ARBUTHNOT S., NAISMITH S. & DANSO S. (2016). Annotated reference corpus of Scottish Gaelic (ARCOSG). http ://dx.doi.org/10.7488/ds/1411.

LIN C.-C., AMMAR W., DYER C. & LEVIN L. (2015). Unsupervised pos induction with word embeddings. *arXiv preprint arXiv :1503.06760*.

LING W., DYER C., BLACK A. & TRANCOSO I. (2015). Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1299–1304.

9. http ://www.faclair.info/

MAOLALAIGH R. Ó. (2013). Corpas na Gàidhlig and singular nouns with the numerals 'three' to 'ten' in Scottish Gaelic. *Scottish Cultural Review of Language and Literature*, **19**, 113–142.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, p. 3111–3119.

MIKOLOV T., YIH W. & ZWEIG G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT-2013)*, p. 746–751.

MILLER G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, **38**(11), 39–41.

PATTON C. (2016). Review of Am Faclair Beag online Gaelic-English dictionary. *Language Documentation and Conservation*, **10**, 155–163.

QIU L., CAO Y. & NIE Z. (2014). Learning word representation considering proximity and ambiguity.

RONG X. (2014). word2vec parameter learning explained. *arXiv preprint arXiv :1411.2738*.

SANTOS C. D. & ZADROZNY B. (2014). Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, p. 1818–1826.

SCANNELL K. P. (2007). The crúbadán project : Corpus building for under-resourced languages. In *Building and Exploring Web Corpora : Proceedings of the 3rd Web as Corpus Workshop*, volume 4, p. 5–15.

SCHNABEL T., LABUTOV I., MIMNO D. & JOACHIMS T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 298–307.

SQA (2009). *Gaelic Orthographic Conventions*. Scottish Qualifications Authority.